engineering proceedings

Designing an ASR corpus for Albanian Language

Amarildo Rista [1], Arbana Kadriu [2]

[1] Aleksandër Moisiu University of Durrës, The Faculty of Information Technology, Durrës, Albania; amarildorista@uamd.edu.al.
[2] South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia; a.kadriu@seeu.edu.mk.
* Correspondence: amarildorista@uamd.edu.al.
† Presented at the 4th International Electronic Conference on Applied Science, 27 Oct–10 Nov 2023, Online.

Abstract: This paper reports the creation of a corpus for Albanian language that is intended for training and evaluating Automatic Speech Recognition (ASR) systems. The corpus comprises 100 hours of audio recordings taken from 200 audiobooks and covers a wide range of topics with a rich vocabulary. The audio recordings are transcribed manually, strictly verbatim, and listened to carefully several times to ensure accuracy. The corpus was evaluated using various end-to-end models as well as Transformer-based architectures. The evaluation was conducted on both the training and testing sets, with Word Error Rate (WER) and Character Error Rate (CER) being considered as evaluation metrics. The results of the architectures trained with this corpus were compared with the results of the LibriSpeech corpus in English. The best architecture based on end-to-end models yielded 5% WER and 1% CER on the training set and 35% WER and 11% CER on the testing set. The transformer-based architecture yielded great results in the testing set, reaching a WER of 18%.

Keywords: Albanian language, Automatic Speech Recognition, Corpus.

1. Introduction

Automatic speech recognition (ASR) systems are designed to convert spoken language into text using machine [1]. The development of ASR systems relies on corpus-driven techniques, and a well-annotated speech corpus is essential for the development and evaluation of these systems. To create a large, high-quality corpus for research and development purposes, significant effort is required for speech data collection, annotation, validation, organization, and documentation. While many efforts have been made for high-resource languages, where ASR systems have made significant progress in performance levels with the help of large speech corpora and deep learning techniques, the development of ASR systems for low-resource languages has remained at very low levels due to the lack of large corpora. In this paper, we present the AlbanianCorpus, a corpus for the Albanian language that contains 100 hours of speech recordings along with their transcripts. The corpus covers a wide range of topics, including biography, social and political sciences, psychology, religion, economics and business, history, philosophy, and sociology, making it as heterogeneous as possible. During the design of the corpus, we considered the characteristics of the Albanian language as well as attributes of the speaker such as age, gender, accent, speed of utterance, and dialect, which are very important in a corpus [2]. To test the validity of the AlbanianCorpus, we trained various end-to-end based architectures [3, 4], as well as Transformers based architectures [5]. We also compared it with the LibriSpeech corpus of the English language [6]. The remainder of this paper is structured as follows: Section 2 reports corpus construction; Section 3 describes the speech recognition architectures designed for the Albanian language; Section 4 focuses on the evaluation of the corpus and presents experimental results; Section 5 lists the conclusions.

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date

2. Corpus Construction

In this section, we present the methodology we follow to build AlbanianCorpus. During its construction, we have considered the features of the Albanian language as well as attributes of the speaker such as: age, gender, accents, speed of speech and dialect.

2.1.Source Data

To select the audio recordings for inclusion into the AlbanianCorpus we have used 200 audiobooks that cover a wide range of topics such as biography, social and political science, psychology, religion, economics and business, history as well as philosophy and sociology. The speeches are recorded in mp3 format and are in both dialects of the Albanian language (tosk and geg). To build a heterogeneous corpus, we have selected speeches from speakers of different age groups, ranging from 20 to 70 years old.

2.2. Creation of Audio and Text files

The audio recordings selected to inclusion into the corpus have varying lengths, ranging from 1 to 3 hours. Since the acoustic model training requires short sequences, to create audio files, we first cut the selected speech data into short sequences (utterances) that range on average from 2 to 12 seconds. Short sequences minimize errors such as disfluencies, repetitions, and corrections, making the acoustic model training achieve the best performance. The cutting of speech data is done using the Audacity tool. [7]. Each speech data sequence created is converted from mp3 to flac format using a 16-bit linear PCM sample encoding (PCM_S16LE) sampled at 22.05 kHz, and is named with a random four-digit decimal number. In case there is a large silence time between the pronunciation of two words, we have cut them leaving no more than 2 seconds length, making the speech sound more polished and confident. To create the text files, we listen carefully several times to each audio file created, writing strictly verbatim the corresponding transcripts for each file. All transcripts are normalized by converting them into upper-case, removing the punctuation, and expanding common abbreviations [8]. The naming of the text file will be done with the same number as the corresponding audio file.

2.3. Corpus description and organization

After creating all the audio files and text files, we have organized them according to the objectives of this study. The final corpus has a size of 12.3 GB with a total duration of speech data of 100 hours. In Table 1 we have reported the size in hours, the number of utterances, the number of words, and the average length of utterance for AlbanianCorpus and the two subsets created, training set and testing set.

Table 1. AlbanianCorpus and its subsets.

Corpus	Size	Number of utterances	Total words	Avg.length per utterance
AlbanianCorpus	100 h	37 358	848 553	9.53 s
AlbanianCorpus_TrainingSet	80 h	29 215	686 971	9.85 s
AlbanianCorpus_TestingSet	20 h	8 543	161 582	8.42 s

3. Speech Recognition Architectures

To evaluate the effectiveness of AlbanianCorpus, a range of models were utilized that incorporate both end-to-end and transformers architectures. The end-to-end models integrate ResCNN [9] and BiRNN [1] as their core components. ResCNN is a deep residual convolutional neural network used for voice activity detection (VAD) and speech enhancement. It is a joint training method that involves a speech enhancement

module to reduce noise and a VAD module to identify speech or non-speech events in a given audio signal. BiRNN is a type of neural network that splits the neurons of a regular RNN into two directions, one for forward states, and another for backward states. Those two states' outputs are not connected to inputs of the opposite direction states, where the output layer can get information from past (backwards) and future (forward) states simultaneously.

ResCNN is fed with Mel Spectogram features from speech signals and is used to identify pertinent features through skip connections, which facilitate the training of deep neural networks by bypassing certain layers, thereby speeding up training and mitigating issues like degradation and vanishing gradients [10]. The outputs from ResCNN are then processed by BiRNN, which takes advantage of the audio features extracted by ResCNN. BiRNN's bidirectional processing enhances the model's predictive accuracy by using a more comprehensive data set.

In contrast, the Transformers model employed is based on the Wav2Vec2 2.0 architecture [5]. It incorporates a multi-layer CNN (encoder) that converts audio wave inputs into latent speech representations. These representations are then discretized using a quantization module. Each encoder block in this model consists of a 1-D fully convolutional network (1D FCN) with causal convolution, followed by a normalization layer and a RELU activation function. To approximate discrete group selection, Gumbel-Softmax distributions are used [11]. Further, a spectral masking technique is applied to these quantized representations [12], which are then fed into the Transformers. These Transformers enrich the entire speech sequence with additional features, building contextual representations [13]. Given the lengthy nature of the masked, quantized representation sequences, a relative positional encoding is used, allowing the Transformers to recognize the sequence order.

4. Results and Discussions

In this section, we present the experiments and results to support our research. First, we train various architectures based on end-to-end approaches. Initially, we evaluate the AlbanianCorpus on the training set and then on the testing set. We also compare the AlbanianCorpus with the LibriSpeech corpus in terms of WER and CER. Second, we train a Transformers-based architecture with AlbanianCorpus and conduct a comparison of the results achieved by end-to-end architectures and Transformers-based architecture.

4.1. Evaluation of the AlbanianCorpus through the training set.

To evaluate the AlbanianCorpus on the training set, we trained five different architectures. The first architecture has 3 RNN layers and 5 GRU layers; the second has 1 RNN and 4 GRU; the third has 1 RNN and 3 GRU; the fourth has 1 RNN and 2 GRU, and fifth has 2 RNN and 2 GRU. All architectures were executed at the same time, on five separate PCs that have the same GPU, CPU and memory. They were trained on the validation set (whole corpus), which contains 100 hours of speech data with their transcripts. All experiments were done in the same conditions related to hyper-parameters. Table 2 shows the results of WER and CER for each architecture.

Table 2. The results of WER and CER on the training set.

Architecture	WER %	CER %	Training time
3 RNN and 5 GRU	5	1	1216 hours
1 RNN and 4 GRU	6	3	967 hours
1 RNN and 3 GRU	7	5	722 hours
1 RNN and 2 GRU	9	8	502 hours
2 RNN and 2 GRU	11	10	915 hours

We show that the architectures with 3 RNN and 5 GRU outperform the rest of the architectures achieving a WER of 5% and a CER of 1% on the training set. We also show that with increasing the number of layers the performance of ASR can be significantly improved. However, we faced difficulties as we were increasing the number of layers beyond 3 RNN and 5 GRU due to increased training time and the unavailability of computational resources. Taking into account the performance of the model and its training time, we have selected the model with 1 RNN and 3 GRU as the most suitable for the continuation of our experiments.

4.2. Evaluation of the AlbanianCorpus through the testing set

For this purpose, we created a training set and a testing set with a split ratio of 80:20. Which means that 80 hours are used for training and 20 hours are used for testing. The selection of data was done randomly. In Figure 1, we reported the results of experiments for both WER and CER related to the number of epochs.

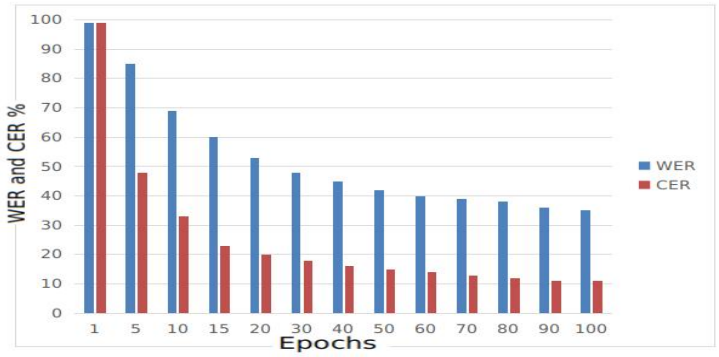
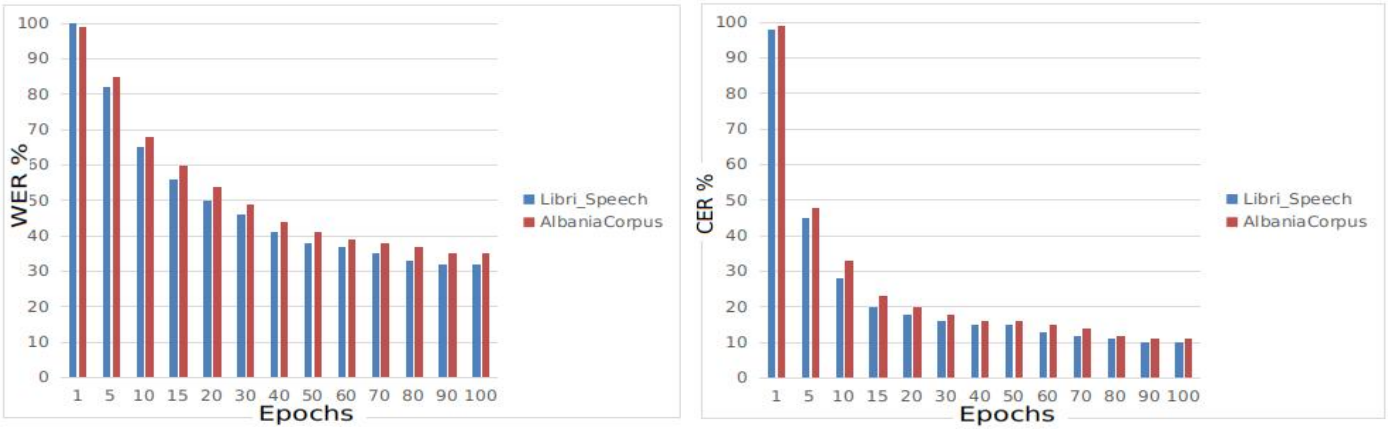


Figure 1. The performance of AlbanianCorpus on the testing set in terms of WER and CER.

As shown in Figure 1, the AlbanianCorpus has yielded a CER of 11% and a WER of 35% on the testing set. The curves for both WER and CER become linear after epoch 50.

4.3. Evaluation of AlbanianCorpus in comparison to LibriSpeech.

To compare these two corpora, we trained the architecture with 1 RNN and 3 GRU layers. First, the model was trained with the AlbanianCorpus and then with LibriSpeech. In both cases, we created a training set and a testing set with a split ratio of 80:20 randomly, specifically the training set of 80 hours and the testing set of 20 hours. All hyper-parameters are the same for both cases. In Figure 2a, we have reported the results for WER related to the number of epochs for both corpora, and in Figure 2b we have reported the results for CER. In the case when the model is trained with the AlbanianCorpus , it has yielded 35% WER and 11% CER on its own testing set. While, in the case when the model is trained with the LibriSpeech corpus, it has yielded 32% WER and 9% CER. We show that AlbanianCorpus yields comparable results to the LibriSpeech corpus.



(**a**) (**b**)

Figure 2. The performance of AlbanianCorpus and LibriSpeech corpora. (**a**) WER results; (b) CER results.

4.4. Evaluation of Transformers architecture using AlbanianCorpus.

In this section, we present the experiments and results for the Transformers-based architecture. The model is trained with the training set and tested with the testing set with a split ratio of 80:20, which means that 80 hours are used for training and 20 hours are used for testing. Splitting of data on both training and a testing subset is done randomly. In Table 3, experimental results are reported. Referring to the WER parameter which is the main indicator of the performance of an ASR system, it has reached 18 %. This result is impressive for the ASR in the Albanian language. Also, from the experimental results, it is noticed that with Transformers the model converges quickly and the training time reaches up to 48 hours. Up to epoch number 10, we received the minimum of WER. After epoch 10, the results undergo negligible change.

Table 3. Experimental results of Transformers-based architecture.

Training Loss	Epoch	Step	Validatio Loss	WER
5.16	0.1	200	29.123	0.9707
0.6853	1	2000	0.3244	0.2906
0.6154	2	4000	0.2861	0.2424
0.5299	4	8000	0.2632	0.2081
0.4563	6	12000	0.2604	0.1997
0.419	8	16000	0.2789	0.1909
0.3758	10	20000	0.2893	0.1863

5. Conclusions

In this paper, we introduced the AlbanianCorpus, a corpus for the Albanian language that contains 100 hours of transcribed audio recordings and covers a wide range of topics with a rich vocabulary. We evaluated the AlbanianCorpus using various end-to-end models as well as Transformer-based architectures. We observed that the model with 1 layer RNN and 3 layers GRU is the most suitable architecture, achieving 5% WER and 1% CER on the training set. To assess the validity of our corpus, we compared it with the LibriSpeech corpus in the English language. Both corpora have been split into training sets and testing sets. The trained AM of each corpus has been evaluated using the testing sets of both corpora. The AM trained with AlbanianCorpus has yielded 35% WER and 11% CER on the testing set, while LibriSpeech has yielded

32% WER and 10% CER. Lastly, we evaluated our corpus by training an architecture based on Transformers. With this architecture, we achieved great results by obtaining a WER of 18% in the testing set. This study introduces a noble contribution for the Albanian language, which is expected to accelerate research within the ASR domain.

Author Contributions: Conceptualization, A.R. and A.K.; methodology, A.K.; corpus design, AR; investigation, A.R and AK.; resources, A.R and A.K.; writing—original draft preparation, A.R; models development, AK; writing—review and editing, A.K.; supervision, A.K. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable
Informed Consent Statement: Not applicable

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kamath, U., Liu, J., & Whitaker, J. (2019). Deep learning for NLP and speech recognition (Vol. 84). Cham, Switzerland: Springer.
2. Lee, J., Kim, K., Lee, K., & Chung, M. (2019). Gender, age, and dialect identification for speaker profiling. In 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).
3. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182). PMLR
4. Rista, A., & Kadriu, A. End-to-End Speech Recognition Model Based on Deep Learning for Albanian. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 442-446). IEEE.
5. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.
6. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.
7. Audacity, T. (2017). Audacity. The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from http://audacity. sourceforge. net.
8. Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. Computer speech & language, 15(3), 287-333.
9. Vydana, H. K., & Vuppala, A. K. (2017). Residual neural networks for speech recognition. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 543-547). IEEE.
10. Ribeiro, A. H., Tiels, K., Aguirre, L. A., & Schön, T. (2020, June). Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness. In International Conference on Artificial Intelligence and Statistics (pp. 2370-2380). PMLR
11. Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144.
12. D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recog-nition," arXiv:1904.08779 [eess.AS], Apr. 18 2019.
13. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805, 2018.